Filtering noise from stock related Twitter messages

Gábor I. Nagy and Sándor Kazi

Twitter is a micro-blogging service that is used by millions of people to publish very short messages and broadcast it to their followers. This real-time service allows users to generate, read and discover interesting content. Parallel to its increasing popularity researchers also took interest in the service. Various research topics include text mining, sentiment analysis, topic discovery, social network analysis. The massive user base - measured in millions of people - also made the service valuable for marketeers, who target consumer with advertisements promoting a particular website or service. An increasing volume of these messages hold little or no information regarding research of regular user content, and this commercial noise bias research and erode the performance of data mining algorithms that extract topics or model sentiment from user generated content. To avoid performance degradation a distinction should be made weather a message is relevant in the context of research or it is just spam.

To tap into the social conversation Twitter allows members to use its API. A prevalent use-case in a data mining is that researchers use the Twitter Stream API. This API helps to gather all the tweets real-time that contain particular terms, stocks (referred to as symbols), hashtags or user entities. For example if a researcher wishes to follow the conversation regarding stocks from S&P500, one can set up the query containing only the symbols of interest. The stream will inevitably contain tweets with a commercial incentive.

In this paper we explore methods for filtering relevant content from Twitter messages related to stocks, indexes and currencies represented by their symbols in tweets. We use the Twitter Streaming API to gather messages real-time for 249 assets. Majority of related work rely on crowdsourced, annotated data to filter relevant messages, user characteristics [1] [2] or textual features [3]. Our approach is comprised of three steps:

1. Label a number of spammers and non spammers and build a classifier to rank users. Features of the classifier include posting behaviour, vocabulary and general characteristics of the user (eg. number of statuses so far, number of friends, number of followers, account age, etc.).

2. Use classifier and rank unseen users. Annotate messages from users with high probability of spamming behaviour. Ranking users and tweets this way helps annotation.

3. Build classifier on messages textual features and evaluate performance.

Evaluation is based on standard accuracy measures (accuracy, precision, recall, F-measure) as well as AUC-ROC. Preliminary results show that ensemble methods, such as Gradient Boosting Decision Tree Classifiers (GBDT) [4] and Random Forest Classifiers (RF) [5] work best on the annotated dataset. The best GBDT model has the following error characteristics based on 5-fold cross validation: mean accuracy 81.6%, mean recall 86.6%, mean f1-measure 82.7% and a mean AUC score of 91.1%. The main characteristics of a spammer is that it uses a lot of different hashtags and stock symbols in its messages with large number of URLs. The account age is found to be less important which is in accordance with the findings in [5], that a lot of spamming accounts are compromised ones. Findings are in line with results in related work experiments. Further refinement of attributes and classifier parameters are needed, with a larger number of annotated users for better classification.

Acknowledgements

This work is supported by the grant: FUTURICT, TÁMOP-4.2.2.C-11/1/KONV, Financial Systems subproject.

References

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter, *CEAS 2010 - Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, July 13-14, 2010, Redmond, Washington, US.

[2] M. McCord, M. Chuah. Spam Detection on Twitter Using Traditional Classifiers, *Autonomic and Trusted Computing*, Lecture Notes in Computer Science Volume 6906, 2011, 175-186.

[3] Juan Martinez-Romo, Lourdes Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, Volume 40, Issue 8, 15 June 2013, 2992-3000.

[4] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, Volume 29, Number 5 (2001), 1189-1232.

[5] Leo Breiman. Random Forests. *Machine Learning*, Volume 45, Issue 1, 5-32.